4.0 VARIABILITY IN CONTEXT

EPA manages the regulation of WET in the same way it manages the regulation of chemical-specific pollutants in order to determine reasonable potential (RP), derive permit limits, determine data quality control, and evaluate self-monitoring data. Many similarities between chemical-specific toxicant and WET controls can be found in the TSD (USEPA 1991a). Determining RP in both cases uses many of the same strategies. Permit limit derivation makes similar exposure assumptions and relies on nearly identical toxicological data bases.

Considering a value other than the best analytical estimate as a measure for WET or for specific chemical analytes is inappropriate. All analytical results, in either chemical-specific analyses or WET tests, incorporate some estimated range of uncertainty. While infrequently discussed for chemical methods, uncertainty does play a role in the meaning of analytical results. One end of the confidence interval likely will be less protective of aquatic resources than the other. The derived limit and therefore final reported analytical results become the best estimate of the actual ecological need and assessment of the effect.

Significant debate has occurred over assertions that WET data have too much inherent variability for reliable use in the NPDES program. This debate has engendered considerable evaluation of WET precision. Groups of scientists and individual researchers have repeatedly concluded that currently promulgated WET methods are technically sound and that the observed precision is within the range of precision of other analyses frequently required in NPDES permits (Grothe et al. 1996). The findings of some of the significant sources of these conclusions are summarized below.

4.1 Society of Environmental Toxicology and Chemistry Pellston WET Workshop

The 1995 Society of Environmental Toxicology and Chemistry (SETAC) Pellston Workshop on Whole Effluent Toxicity convened 47 experts in the discipline to assess applied methods and their application in the regulatory process. Representation at the workshop was intentionally balanced among government, business, and academic participants. These scientists published consensus conclusions and recommendations, including the following.

4.1.1 General Conclusions and Recommendations

Grothe et al. (1996) state *"Existing WET testing methods (USEPA 1985, USEPA 1988, USEPA 1989) are technically sound, but certain modifications would improve endpoint interpretation. Such changes involve implementing improvements to currently used statistical procedures, establishing acceptable limits for MSD values, and adding confidence limits to WET test endpoints."*

"A number of problems with WET tests are caused by misapplication of the tests, misinterpretation of the data, lack of competence of the laboratories conducting WET testing, poor condition/health of test organisms, and lack of training of laboratory personnel, regulators, and permittees. More widespread use of WET related guidance provided in USEPA's TSD (1991a) would help alleviate some of these problems. In addition, an effective QA/QC program will improve data quality and reduce test variability."

"Increase training opportunities for regulators and permittees to improve the implementation of WET objectives and to promote national consistency in permitting and compliance issues."

"Implement a broadly based and standardized QA/QC program to improve WET testing performance and data quality."

"Quantify the 'confidence' around test endpoints to improve interpretation of WET test results. Specific statistical methods that could improve precision are presented in Chapter 3 of this document and processes to reduce variability are discussed in Chapter 5. In addition, WET tests should be performed using a dilution series of exposure concentrations to establish a dose-response relationship."

4.1.2 Conclusions about Data Precision

Ausley (1996) compared CVs of chemical analyses and aquatic toxicity tests conducted by North Carolina NPDES permittees. Ausley found that CVs of reported values for chemical analytes (including metals, organic analytes, and non-metal inorganic analytes) ranged from 11.8 percent to 291.7 percent. Coefficients of variation for toxicity parameters (acute and chronic *Ceriodaphnia dubia*, acute and chronic *Pimephales promelas*, acute *Daphnia pulex*, and acute *Mysidopsis bahia*) ranged from 14.8 percent to 67.6 percent. From this review, he concluded that "*the precision of toxicity analyses is within the range of that being reported for commonly analyzed and regulated chemical parameters.*" Ausley highlighted the difficulty in comparing precision estimates of chemical analytes and WET analyses (particularly NOECs), noting that while chemical precision is often determined well above analytical detection, WET precision is often based on the minimum detection level. An assumption that WET precision will vary among toxicants is also logical. To establish "inherent variability," considering toxicants that cause minimal variability in the analysis may be appropriate. The high coefficients of variation for some chemical parameters reported by Ausley reflect the fact that, in practice, analytical precision can vary widely in individual studies in which the effects of a single (or a few) poorly operating laboratory can adversely affect precision estimates. In practice, this kind of data must be screened for quality prior to use to evaluate self-monitoring data or estimates of overall method quality.

Ausley's results closely approximate analytical precision of chemical analytes referenced in the TSD (USEPA 1991a, Chapter 1.2). The CVs for metals (aluminum, cadmium, chromium, copper, iron, lead, manganese, mercury, silver, and zinc) ranged from 18 percent to 129 percent at the low end of the measurement detection range. Between-laboratory CVs for organic analytes ranged from greater than 12 percent to 91 percent. The CVs for non-metal analytes (alkalinity, residual chlorine, ammonia nitrogen, Kjeldahl nitrogen, nitrate nitrogen, total phosphorus, biological oxygen demand, chemical oxygen demand, and total organic carbon) ranged from 4.6 percent to 70 percent in between-laboratory studies of precision.

Burton et al. (1996) concluded that "*USEPA-published methods are functional and appropriate in the context of effluent toxicity control programs.*" They recommended developing limits on within-test variability, a quality assurance and audit program, and guidance for permittee procurement of WET analytical services.

Denton and Norberg-King (1996) cited various studies that favorably compare WET methods with chemical analytical methods (Grothe and Kimerle 1985, Rue et al. 1988, Morrison et al. 1989, Grothe et al. 1990). They proposed that improvements in test result consistency could be accomplished by limiting the range of within-test variability through controls of upper and lower statistical power (e.g., limits on test MSD). Three practices to control within-test variability most effectively are (1) controlling within-test sensitivity, (2) following well-defined test methods, and (3) maintaining communication within the regulatory community. For example, the permittee and regulatory authorities should discuss any facility-specific issues to fully characterize the appropriate permit conditions.

4.2 Water Environment Research Foundation Study

Another publication, "*Whole Effluent Toxicity Testing Program: Evaluation of Practices and Implementation*" (DeGraeve et al. 1998), presents the results of a survey of publicly owned treatment works and State regulatory programs about WET issues. The Water Environment Research Foundation (WERF) sponsored this study. Conclusions by DeGraeve et al. (1998) include the following:

> "*The project team believes that the results demonstrate that the test methods can be routinely completed successfully by well-trained, competent WET testing laboratories and that the results, considered collectively, suggest that the test methods that are being used to measure WET are technically sound."*
>
> "*There is a need for better training/guidance in WET-related issues for both the regulatory staff responsible for implementing WET requirements and for permittees responsible for meeting WET limits*."

DeGraeve et al. (1998) considered the conclusions of the SETAC Pellston WET publication concurring that between-laboratory CV values of toxicity test methods were low, training of regulatory and permittee staff is needed nationally, and strengthened quality assurance (QA)/quality control (QC) practices could improve performance of analyses. Unlike the SETAC Pellston WET conclusions, they found that there are enough laboratories to meet the current market demand for analyses. Like the SETAC effort, DeGraeve et al. (1998) concluded that a national center of expertise on WET issues would be beneficial to provide guidance to regulatory agencies, permittees, and laboratories.

WERF also funded a project entitled "*Whole Effluent Toxicity Testing Methods: Accounting for Variance*" (Warren-Hicks et al. 1999). This study compared within- and between-laboratory results of reference toxicant test variation as measures of reproducibility and comparability, respectively. The authors concluded that some laboratories could consistently reproduce test results, while others could not and inferred that test precision is a factor of laboratory experience and not inherent methodological weakness. The authors recommended that national studies be conducted to evaluate within- and between-laboratory precision of promulgated WET test methods. (EPA has already initiated this study.) They also recommended that additional test acceptability criteria (TAC), such as upper and lower bounds of MSD, be established and incorporated in the NPDES process. The latter recommendation corroborates other researchers' recommendations discussed above.

4.3 Minimizing Variability by Adhering to WET Toxicity Test Methods

Specific factors that affect variability in WET analyses have been described in several papers (Burton et al. 1996, Ausley 1996, Erickson et al. 1998, Davis et al. 1998). The most important initial consideration in developing precise data is a laboratory's experience and success in performing a specific analysis. Most critical reviews of WET data precision emphasize this initial consideration. Experienced professionals most likely will be able to develop the most consistent and reliable information and can interpret anomalous conditions in the testing or results.

An additional factor in considering WET test method variability is whether the prescribed methods (e.g., the EPA toxicity test methods promulgated in 40 CFR Part 136) are being followed appropriately (see Chapter 5). If tests are submitted that do not meet specified TAC or are produced when laboratory QA testing indicates analyses are beyond control limits, these results should not be used in the NPDES process. Tests performed on effluent samples that have not met required temperature maxima or holding times should not be considered for regulatory purposes. Rigorous QA practices are critical to the success of any analytical program. Both the regulatory authority and permittee should strive to ensure that such practices are in place

for any program developing WET data, whether by national laboratory accreditation, State regulatory certification, direct permittee oversight, or specific contractual agreement with the laboratory.

Comparisons of WET method precision with analytes commonly limited in NPDES permits clearly demonstrate that the promulgated WET methods are within the range of variability experienced in other analyses. Several researchers also noted clear indications that method performance improves when prescribed methods are followed closely by experienced analysts (Grothe et al. 1996, DeGraeve et al. 1998).

A review of WET test results confirms that imprecise WET data are being reported. As with any analytical technique, inexperienced individuals can perform analyses incorrectly or fail to follow appropriate methods and quality assurance practices. Using the training that is available for these methods and quality assurance techniques referenced by this document will help ensure that data of maximum reliability are used and that sound decisions are made based on those results. The Western Coalition of Arid States conducted a study in 1997 (Moore et al. 2000), which reported the results of 16 tests with a non-toxic test sample using the *Ceriodaphnia dubia* chronic test. These results indicated that 43 percent of the tests showed toxicity. EPA is in the process of reviewing the paper and the raw data.

Persons interested in WET issues may consult another source of information developed by the SETAC Whole Effluent Toxicity Expert Advisory Panels. This group, established under a cooperative agreement with EPA, provides scientific opinion and training on WET technical issues. This information is available on the Internet at the SETAC web site, **http://www.setac.org**. Appendix D contains frequently asked questions with answers prepared by the SETAC WET Expert Advisory Panels. The expert panels have identified and discussed various factors that affect WET variability.

4.4 Conclusion

When the variability of WET analyses is viewed in the context of the NPDES program, these techniques produce data that are as precise as those from chemical analyses. As with any other analytical system, lack of experience in performing the analyses, adherence to prescribed QA practices, or good laboratory practices will reduce the precision of the results. Studies of these factors by independent researchers from both the regulatory and regulated communities support these conclusions. While examples of poor-quality, highly variable results from chemical analyses have also been publicized, these results are frequently influenced by the shortcomings mentioned above. Permittees that must generate and use WET data should become well-educated in data quality interpretation, and permittees should require that QC practices be followed by laboratories generating the data. Various sources of information presented in this chapter should assist permittees, testing laboratories, and regulatory authorities with this education process. Examples of practices that can further reduce the imprecision of analyses are also discussed in Chapters 5 and 6 of this document. Additional refinements of TAC can likewise improve test power to detect effects (or the lack thereof) and increase the statistical confidence in results.